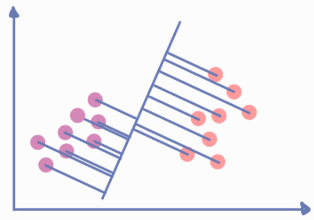
Fisher's linear discriminant (LDA - Linear Discriminant Analysis)



This model does classification in terms of dimensionality reduction. LDA uses the information from all the features to create a new axis and projects the data over it in a way that it maximizes the separation of the classes.

The simplest measure of the separation of classes when projected onto w, is the separation of the projected class means.

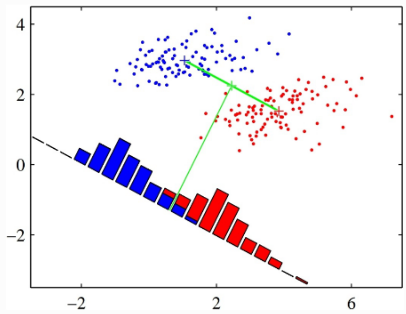
$$\text{maximize } J(w) = w^T (m_2 - m_1) \qquad m_1 = \frac{1}{N_1} \sum_{n|x_n \in C_1} x_n \qquad m_2 = \frac{1}{N_2} \sum_{n|x_n \in C_2} x_n$$

FOR 2 CLASSES

this expression can be made arbitrarily large simply by increasing the magnitude of w; constrain w to unit length → $|w| = 1$

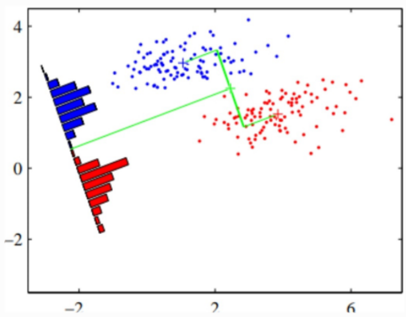
Consider the case in which we have only two classes; the concept is the same for a multiclass problem:

① compute the mean for the two distributions;
② plot the line connecting the two means;
③ project all the points onto this line;
④ choose a threshold and classify the points on its left as one class and the points on its right as the other class.



$$w \propto (m_2 - m_1)$$

We use the mean but we are not using the covariance. Add information on the covariance allowing the line also to rotate



$$w \propto S_W^{-1} (m_2 - m_1)$$

Rotation matrix

$$w_0 = w^T m$$

with m = global mean of all the dataset

Once we compute weight matrix and w_0
↓
PREDICTION

Fisher's linear discriminant is given by the function $y = w^T x$ and the classification of new instances is given by $y > -w_0$

PCA → reduces dimensions by focusing on the features with the most variation
LDA → focuses on maximizing the separability among classes

Multiple classes:

$$y = W^T x$$

$$J(W) = Tr\{(W S_W W^T)^{-1} (W S_B W^T)\}$$

...